Recommendation Challenge : HackerRank

Aim: Recommend 10 challenge to a particular Hacker\_ID given it should be unsolved by the user.

Approach:

(1) Given two csv files containg Submission record and user records. Starting with data analysis, i found in submission record under domain and sub-domain variables more than 50% data where missing so i dropped them out.

(2) Further created a pivot table combining both tables. Hacker\_id was taken as index and challenge\_id as columns.

(3) The resultant table each row shows performance of each hacker in different compettions and challenges.

(4) Now in order to suggest some problems to user we must know new problems which other users have tried we need to cluster similar users. Doing so the problem subset reduces.

(5) After applying Kmeans i got 5 clusters. And applied PCA to reduce the dimesion(=2). And added the 'x' and 'y' co-ordinates of cluster points as a new variable giving more insight of how close they are.

(6) Then finally i seperated out the clusters and passed them to my recommendation function.

(7) Now in this function i calculated the similarity amongst the user among the cluster ussing hamming distance becuase i was using both categorical and numerical features.

(8) After calculating the hamming distance of a hacker in cluster with rest hackers i sorted the distance and selected the challenges solved by the close ones.

(9)Then i found out all the top 100 hackers who solved /attempted those problems from original table with other features also including the language they used.

(10) Then for each hacker i seperated them out from original table and found the challenges they have solved which has not been solved by given user and thus recommending them to the hacker.